

*“Los números son el principio de las cosas”
Pitágoras*

3. TIPOS DE DATOS

No es raro que aún los que están acostumbrados a manejar datos ecológicos desconozcan su correcta denominación. Sin embargo, existen distintas categorías y varias clasificaciones (véanse a Sokal y Sneath, 1963; Clifford y Stephenson, 1975; Boesch, 1977; Orlóci, 1978; Crisci y López Armengol, 1983) cuyo conocimiento permite hacer más eficiente su obtención primaria, convertir un tipo en otro, explotar al máximo su contenido de información, conocer las limitaciones que cada tipo encierra y determinar las exigencias que a ellos pueden aplicarse (Tabla 3.1).

Las clasificaciones y categorías existentes están referidas a distintas manifestaciones del quehacer científico, por lo que de ellas tomaremos solo aquellas que son de uso más común en ecología (Tabla 3.2). Para la taxonomía numérica el interesado puede consultar a Sokal y Sneath (1963) o Dunn y Everitt (1982) que tratan en detalle los tipos de datos de esta disciplina. De cualquier forma, sea cualfuere el tipo de dato, su clasificación procede a partir del número de estados que alcanza (doble o múltiple), su naturaleza (cualitativos y cuantitativos) y su escala de medición.

Tabla 3.1. Clasificación general de los tipos de datos.

Número de estados	Naturaleza	Forma de estado
DOBLE ESTADO	Cualitativos	Presencia-Ausencia Estados excluyentes
MULTIESTADO	Cualitativos	Ordenado Desordenado
	Cuantitativos	Continuos Discontinuos
	Semicuantitativos	Rangos Escalas

Tabla 3.2. Tipos de datos de mayor aplicación ecológica y ejemplos.

Tipo de dato	Naturaleza	Posibles estados	Ejemplos de parámetros o situaciones ecológicas
DOBLE ESTADO	Cualitativo	Presente Ausente	Composición cualitativa de especies
MULTIESTADO	Cualitativo ordenado	Abundante Moderado Escaso	Ordenamiento cualitativo de parámetros ecológicos
	Cuantitativo discontinuo	0 a α	Número de taxones, número de individuos
	Cuantitativo continuo	0 a α	Abundancia, biomasa, cobertura, diversidad
	Semicuantitativo	1 a α	Asignación de rangos y escalas de abundancia

Datos cualitativos

Datos de doble estado: presencia-ausencia.- Los datos de doble estado, siempre cualitativos, son los llamados *binarios* o predicados dicotómicos (Crisci y López Armengol, 1983). El tipo de mayor aplicación ecológica es el de *presencia-ausencia*, ejemplificado cuando en nuestras estaciones, unidades muestrales o intervalos de tiempo, evaluamos el número de especies y dónde éstas están o no presentes. Poseen solamente dos estados de carácter, en este caso presencia o ausencia de las especies, de ahí su denominación de binarios.

En el muestreo ecológico el dato de presencia-ausencia identifica si la especie aparece o no en una estación determinada, por lo que podemos expresar la presencia con un signo “+” ó una cruz y la ausencia con un signo “-” ó un espacio en blanco. Sin embargo, como la clasificación exige una expresión cuantitativa del dato primario para los cálculos, el dato cualitativo de doble estado debe ser codificado (Fig. 3.1). Por ello, numéricamente se expresa como «1» la presencia y como “0” la ausencia, donde el número desempeña sólo una función de rótulo o marca de identificación para facilitar el posterior tratamiento cuantitativo (Crisci y López Armengol, 1983).

Especies	Estaciones				
	1	2	3	4	5
A	+	+	+		
B	+		+	+	+
C	+		+		+
D	+	+	+	+	+
E	+	+	+	+	+

Especies	Estaciones				
	1	2	3	4	5
A	1	1	1	0	0
B	1	0	1	1	1
C	1	0	1	0	1
D	1	1	1	1	1
E	1	1	1	1	1

Figura 3.1. Matriz original de datos de presencia ausencia y de datos codificados.

A este tipo de dato se denomina *asimétrico* pues los estados que se asignan no tienen el mismo peso a la hora de agrupar (Kaufman y Rousseeuw, 1990). Ello se debe a que cuando decimos que dos especies comparten una estación implica que tienen algo en común pero no puede decirse exactamente lo mismo si no están presentes, por lo que los “ceros” pueden ser menos importantes que los “unos”. Contrariamente, el otro tipo de dato cualitativo de doble estado: el de *estados excluyentes* se identifica como *simétrico* pues ambas entidades tienen el mismo peso. Un ejemplo conocido es el sexo donde si adjudicamos un código de “1” a los machos y “0” a las hembras ambos son igualmente importantes a la hora de agrupar, aunque el dato de estados excluyentes no es de mayor aplicación en la ecología estructural.

Al emplear datos binarios en el análisis normal deseamos conocer el grado de parecido entre las listas de especies en las localidades muestreadas. Como sólo se cuenta con la presencia-ausencia como atributo puede ocurrir que dos especies tengan una afinidad cualitativa muy alta por poseer similar distribución en las dos localidades, aún cuando sus dominancias o abundancias sean bien diferentes. Por ello, Pielou (1977) plantea que el dato de presencia-ausencia puede conducir a una

sobreestimación desproporcionada de las especies escasas en relación con lo que representan sus valores reales de abundancia, que cuantitativamente serían muy pequeños o despreciables. En el análisis inverso, donde nos interesa conocer cuál es el grado de concurrencia de las especies, o sea cuáles comparten determinadas localidades puede ocurrir que dos especies aparezcan siempre juntas -lo que también brindaría una afinidad cualitativa máxima entre ellas-, pero por la acción de determinados factores bióticos o abióticos que promuevan su incremento en algunos hábitats o su disminución en otros, se presenten con abundancias o dominancias diferentes.

Aunque lo anteriormente señalado son desventajas reales de los datos binarios, algunos trabajos los señalan como más ventajosos que los cuantitativos al no estar sujetos, en la misma medida, a la influencia de las variaciones estacionales o locales, o a los errores propios del muestreo cuantitativo. Independientemente de esto, los datos binarios suelen ser la única alternativa en algunos campos como la Biogeografía, donde solo se cuenta con listas de especies por regiones. Además, si la matriz original posee un alto porcentaje de ceros, o sea, el ambiente estudiado es pobre en especies, la información que aporta la cuantificación puede ser mínima.

Por otra parte, siempre que se estudie un gradiente ecológico de cambios y la variación a estudiar tenga un reflejo en la composición cualitativa de la comunidad, es preferible esta alternativa. Tal es el caso de la zonación de las especies de moluscos en el litoral rocoso (Herrera *et al.*, 1987) que muestra un patrón de variación del mar hacia la costa, como reflejo de adaptaciones particulares a los tensores propios de este ambiente (Tabla 3.3). Especies como *Drupa nodulosa*, *Astraea tuber* y *Chiton marmoratus*, por mencionar algunas, tipifican los primeros pisos litorales como componentes de la región infralitoral; *Acanthopleura granulata* y *Acmaea cubensis* la región mesolitoral; mientras que *Cenchritis muricatus*, *Littorina lineolata* y *Echininus nodulosus* caracterizan el extremo ambiental supralitoral. Nótese como la propia tendencia de aparición de los datos en la matriz, aún sin ordenar, sugiere la distribución por horizontes litorales.

Para brindar otro ejemplo de cómo la tendencia de grupos puede observarse en la propia matriz de datos cualitativos tomamos la información del estudio de la composición y diversidad de las pesquerías en ocho sitios de desembarco de la Bahía de Samaná, República Dominicana (Sang *et al.*, 1997) y reordenamos sus datos con un sentido ecológico-geográfico, del estuario al océano. La Tabla 3.4 muestra cambios paulatinos en las familias representadas en los diferentes sitios de desembarco, que varían según su posición geográfica respecto a los diferentes ambientes de pesca adyacentes: estuario, manglares, pastos marinos y arrecifes coralinos.

En las áreas del NO de la Bahía (Sánchez) se registran las mayores capturas de camarones peneidos y una ictiofauna típicamente estuarina, con especies demersales pertenecientes a las familias Centropomidae y Mugilidae o pelágicas de Engraulidae. En la medida que abandonamos la influencia del estuario y comienza el desarrollo de pastos marinos y arrecifes coralinos, como ocurre en Las Pascualas, Los Cacaos (al N y centro de la bahía), Sabana la Mar y Miches (al SE casi fuera de la bahía) aumenta la representación de familias demersales neríticas arrecifales como Serranidae, Holocentridae, Balistidae o Scaridae aunque por la situación de gradiente también se registran algunos representantes estuarinos y/o típicamente pelágicos. En Las Galeras y Las Terrenas (fuera

Tabla 3.3. Distribución cualitativa de treinta y cinco especies de moluscos en veinte unidades muestrales del litoral rocoso de Puerto Escondido, (NO de Cuba) ubicadas del mar a la tierra (según Herrera *et al.*, 1987).

Especies	Unidades muestrales																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
<i>Drupa nodulosa</i>	X																			
<i>Astraea tuber</i>		X																		
<i>Chiton marmoratus</i>	X				X	X														
<i>Acantochitona astriger</i>			X	X	X															
<i>Leucozonia ocellata</i>				X																
<i>Thais rustica</i>							X													
<i>Hemitoma octoradiata</i>	X	X	X	X																
<i>Tricolia adamsi</i>		X	X		X															
<i>Fissurella barbouri</i>	X	X	X			X														
<i>Fissurella nodosa</i>	X		X	X			X													
<i>Pinctada radiata</i>	X			X																
<i>Planaxis lineatus</i>		X			X															
<i>Diodora viridula</i>	X	X				X														
<i>Columbella mercatoria</i>	X	X	X		X		X			X										
<i>Fissurella angusta</i>	X			X								X								
<i>Fissurella rosea</i>		X				X		X												
<i>Thais deltoidea</i>							X	X												
<i>Fisurella barbadensis</i>				X			X	X			X	X								
<i>Petalococonchus irregularis</i>					X	X	X	X		X										
<i>Ceratozona squalida</i>				X			X	X	X	X										
<i>Acmaea leucopleura</i>		X		X	X			X	X	X										
<i>Cittarium pica</i>				X			X			X	X									
<i>Littorina meleagris</i>			X	X	X	X	X	X	X	X	X									
<i>Brachidontes exustus</i>	X		X	X	X		X	X	X	X	X			X						
<i>Isognomum alatus</i>	X			X		X	X	X												
<i>Acmaea jamaicensis</i>	X	X	X	X	X	X	X	X			X	X								
<i>Purpura patula</i>								X												
<i>Acanthopleura granulata</i>										X										
<i>Littorina ziczac</i>								X												
<i>Acmaea cubensis</i>											X									
<i>Littorina angustior</i>												X	X		X					
<i>Nodilittorina tuberculata</i>													X	X		X				
<i>Cenchritis muricatus</i>													X		X	X	X			
<i>Littorina lineolata</i>													X		X		X			
<i>Echininus nodulosus</i>													X	X	X		X	X	X	

de la bahía y al N de la Península), que representan el extremo del gradiente hacia el océano, algunas de las anteriores familias está ausentes, aparecen otras del complejo de pastos marinos- arrecifes coralinos (indicando su relación con estos ecosistemas) y de manera exclusiva las familias Coryphaenidae e Istiophoridae, de hábitos pelágicos, en concordancia con su ubicación en las cercanías de mar abierto.

De cualquier forma, la elección de la variante cualitativa no excluye que debamos tener en cuenta elementos cuantitativos. El número de especies que se registra en una comunidad guarda una estrecha relación con el número de individuos muestreados, de forma tal que sólo con un tamaño de muestra

Tabla 3.4. Matriz de presencia-ausencia de las principales familias de crustáceos y peces de las pesquerías de Samaná en siete sitios de desembarco, reordenados en un gradiente del estuario al océano, a partir de datos de Sang *et al.* (1997). Las letras indican los sitios de desembarco: SZ: Sánchez, M: Miches, LC: Los Cacaos, SM: Sabana de la Mar, LP: Las Pascualas, LG: Las Galeras y LT: Las Terrenas.

Familias	Ambientes: Estuario-Manglares-Pastos-Arrecifes-Océano						
	Sitios de desembarco						
	SZ	M	LC	SM	LP	LG	LT
Pristigasteridae	X						
Ophidiidae	X						
Cynoglosidae	X						
Elopidae	X						
Engraulidae	X	X					
Stromateidae	X	X		X			
Trichiuridae	X	X			X		
Clupeidae	X	X	X		X		
Penaidae	X	X		X	X		
Synodontidae	X		X		X		
Mugilidae	X	X		X	X		
Centropomidae	X	X		X			
Ephippidae	X	X		X	X		
Gerreidae	X	X	X	X	X	X	X
Scianidae	X	X		X	X	X	X
Carangidae	X	X	X	X	X	X	X
Lutjanidae	X	X	X	X	X	X	X
Haemulidae	X	X	X	X	X	X	X
Sphyraenidae	X	X	X	X	X	X	X
Scombridae	X	X	X	X	X	X	X
Polynemidae		X		X	X	X	
Scaridae		X	X	X	X	X	X
Mullidae		X	X	X	X	X	X
Sparidae		X	X	X	X	X	X
Serranidae		X	X	X	X	X	X
Holocentridae		X	X	X	X	X	X
Priacanthidae		X	X	X	X		X
Balistidae		X	X	X	X	X	X
Belonidae			X	X	X	X	X
Palinuridae		X	X	X		X	X
Labridae		X	X	X		X	X
Acanthuridae			X	X		X	X
Ostracidae			X		X	X	
Pomacanthidae				X	X	X	
Coryphaenidae						X	X

elevado el total de especies aparece adecuadamente representado. Quiere esto decir, por ejemplo, que no es conveniente para propósitos clasificatorios comparar dos localidades cuyas listas provengan en un caso de una muestra de 300 individuos y en la otra de 50, a causa de serias diferencias en el esfuerzo de muestreo. En el segundo caso, el requisito de tamaño de muestra que hace que la lista de especies obtenida sea representativa de la localidad estudiada, probablemente no se cumpla.

El tamaño mínimo de muestra para obtener una representación adecuada de la composición específica se estima haciendo un gráfico acumulativo del número de especies contra el número de individuos o unidades muestrales. La curva acumulativa, muy variable al inicio, va tendiendo a una estabilización al crecer el tamaño de la muestra hasta llegar a su nivelación horizontal, punto en el cual se considera que el número de individuos o unidades muestrales es suficiente para estimar adecuadamente al número real de especies.

No obstante, una comunidad puede tener una abundancia tan baja que sea prácticamente imposible obtener un tamaño de muestra grande. Por ejemplo, en el arrecife del Rincón de Guanabo, en la costa Norte de la Habana, Cuba, 30 unidades muestrales de 1 m² son suficientes para coleccionar 400 individuos, representados por 20 especies de gorgonáceos en los pavimentos rocosos de 10 m; mientras que en la laguna arrecifal, un esfuerzo de muestreo igual en términos de unidades muestrales, arroja solamente 21 individuos y 3 especies (Herrera *et al.*, 1997). En tales casos la representación de especies es adecuada para fines clasificatorios pues la comunidad está condicionada por un conjunto de tensores (presencia de sedimento, cobertura vegetal, turbidez y batimiento) que no permiten una mayor abundancia y no tiene sentido incrementar el esfuerzo de muestreo.

Datos cualitativos de multiestado ordenado.- Los datos de *multiestado ordenado*, también denominados *de secuencia lógica*, se refieren a datos cualitativos que pueden ser ordenados en una secuencia de magnitud de la cualidad estudiada (Crisci y López Armengol, 1983), pues poseen una jerarquía de formas diferentes que abarcan la variación total del intervalo de entidades. Un ejemplo típico es cuando en una lista de especies vamos más allá de señalar su simple presencia o ausencia y tenemos criterios para subdividirlas en: abundantes, comunes, escasas y raras; dentro de la lista varias especies pueden recibir la misma categoría.

Como se trata en definitiva de un dato cualitativo de presencia-ausencia con un poco más de información, esta ganancia informativa solo puede ser aprovechada si se plasma cuantitativamente, o sea, codificando las especies mediante puntajes del 1 en adelante (Fig. 3.2). En tal caso, los números desempeñan una función ordinal ya que indican la posición de una cualidad en una secuencia de grados (Crisci y López Armengol, 1983). Este ordenamiento en categorías diferentes debe partir de un criterio de magnitud que permita la jerarquización.

Especies	Estaciones				
	1	2	3	4	5
A	Ab	Ab	Co	Au	Ab
B	Co	Co	Ab	Co	Au
C	Es	Ra	Co	Ab	Au
D	Ra	Ra	Co	Co	Au
E	Au	Ra	Ra	Ra	Au

Especies	Estaciones				
	1	2	3	4	5
A	4	4	3	0	4
B	3	3	4	3	0
C	3	1	3	4	0
D	1	1	3	3	0
E	0	1	1	1	0

Figura 3.2. Matriz original de datos de multiestado ordenado y de datos codificados, considerando: Ab: abundante (4), Co: común (3), Es: escasa (2), Ra: rara (1), Au; ausente (0).

Esta codificación de los datos en clases implica que mientras mayor es el número de clase, mayor es la abundancia, pero una misma clase no siempre significa una abundancia igual. Los intervalos entre clases no tienen casi significación y dependiendo de sus límites la agrupación puede estar influida en varias formas (van Tongeren, 1987) razón por la cual no se recomiendan para la clasificación (Boesch, 1977). Este tipo de dato es producto de un muestreo cualitativo donde el componente cuantitativo ha sido manejado de forma grosera, un muestreo cuantitativo donde la precisión de los datos es dudosa, o simplemente una división basada en la experiencia. Por ejemplo, en el arrecife costero al N de La Habana, Cuba, son abundantes los corales *Montastrea cavernosa* y *Siderastrea radians*; *Dichocoenia stokesi* es común; y *Colpophyllia natans* es rara.

A diferencia de los datos de multiestado ordenado, los datos de *multiestado desordenado*, o *sin secuencia lógica* no pueden ser organizados en una secuencia de grados del atributo. Un ejemplo sería la relación de una especie con el sustrato: infaunal, epifaunal y críptica. Evidentemente no habría criterios para codificar estas tres formas como 1, 2 y 3 pues no hay ninguna jerarquización. Esto se resuelve convirtiendo cada estado en un dato de presencia ausencia (por ejemplo: críptica (1), no críptica (0)) pero ello implica una sobreestimación del atributo “relación con el sustrato”, por lo que la subdivisión ordinal de los datos puede ser subjetiva y crear discontinuidades que realmente no existen. Este tipo de dato ha quedado para otras disciplinas como la taxonomía numérica y no es de mayor aplicación ecológica como el anterior que si es posible su ordenación. Aún así, es una forma de hacer más refinada la información que de acuerdo a la práctica más común, nos conducirían, como veremos inmediatamente, al concepto de rango.

Si categorizamos los datos cualitativos hasta aquí vistos, considerando su escala de medición y sus propiedades tendríamos dos grupos, según describe Sharma (1996). Los datos de presencia-ausencia, estados excluyentes y multiestado desordenado que tienen una escala de medición nominal y los números son usados solo para categorizar, son inapropiados para cálculos de estadígrafos como la media y la desviación estándar y solo las estadísticas basadas en conteos como la moda o las distribuciones de frecuencias, son apropiados para ellos. El tipo de dato es, por tanto, *nominal*. En los datos de multiestado ordenado existe una significación de jerarquías pero ésta es simplemente de escala ordinal pues la secuencia de categorías, no representa diferencias iguales del atributo medido. Los estadígrafos válidos para este tipo de datos son la moda, la mediana, las distribuciones de frecuencia y los métodos no paramétricos como la correlación por rangos (Sharma, 1996). El tipo de dato es llamado *ordinal*. Las variables que se miden usando escalas nominales u ordinales se refieren comúnmente como variables *no métricas*.

Datos de rango. - Cuando al dato de multiestado cualitativo ordenado se le asigna un número, en vez de un estado jerárquico nominal (abundante o escaso, por ejemplo), entonces la secuencia de magnitud adquiere un carácter semicuantitativo y hablamos de rangos. El dato de rango gradúa la colección especie por especie y puede partir tanto de un criterio general de abundancia, según ya vimos, como de valores numéricos individuales que permitan diferenciar entre las especies abundantes, la primera, la segunda o la tercera.

Quizás pueda pensarse que si cada especie está caracterizada por un número representativo de su abundancia o proporción no es necesario asignar rangos, pues la clasificación puede proceder a partir de ellos. Esto es cierto, pero la gradación en rangos es una alternativa útil cuando la precisión de los valores es cuestionable por errores en el muestreo, número insuficiente de réplicas o diferencias notables en el esfuerzo, entre otras causas. El uso de rangos puede ser parte de una estrategia de muestreo donde se busque rapidez y se desee obtener un panorama general de la comunidad, con mayor ganancia informativa que con datos de multiestado codificados. En tal caso, el ordenamiento sucesivo de las especies en rangos mantiene el componente cualitativo de la información y aprovecha el cuantitativo que de otra forma se perdería.

La forma más común de asignar los rangos (Fig. 3.3) tomada de la estadística no paramétrica (Siegel, 1985), se basa simplemente en substituir los valores originales por números a partir del uno, de mayor a menor. Ante datos repetidos en la matriz original, como por ejemplo las especies C, D y E en la estación 3 que coinciden en el valor 5, se le asigna a cada una el rango que le corresponde en orden, en este caso 3, 4 y 5 pero el valor final de cada una será el promedio: $(3 + 4 + 5)/3$, o sea 4. Otro tanto ocurre con las especies E y F, en la estación 5.

Especies	Estaciones					Especies	Estaciones				
	1	2	3	4	5		1	2	3	4	5
A	10	56	150	1	45	A	2	1	1	5	1
B	20	50	32	0	20	B	1	2	2	6	2
C	5	12	5	114	19	C	3	3	4	1	3
D	3	10	5	7	5	D	4	4	4	4	4
E	2	9	5	20	1	E	5	5	4	3	5,5
F	1	0	1	32	1	F	6	6	6	2	5,5

Figura 3.3. Matriz original de datos cuantitativos y matriz de rangos.

Este tipo de dato no debe usarse como norma habitual de trabajo, recuérdese que no representa el dato real e incluso valores iguales pueden recibir rangos diferentes o valores muy desiguales pueden coincidir en un mismo rango. Por ello, su empleo debe ser limitado a aquellas medidas de afinidad, que como veremos, se basan precisamente en datos de jerarquía.

Finalmente, vamos a referirnos a los datos derivados del empleo de *escalas de abundancia*, que si bien no son datos de rango en el sentido antes explicado, tienen también la característica de que un valor (del 1 en adelante) sustituye el dato primario; en estos casos siempre originalmente cuantitativo. Una escala de tal tipo (Tabla 3.5), cuya significación ecológica está avalada estadística y prácticamente, la ofrece Frontier (1969). Aunque el empleo de escalas de abundancia tiene como objetivo realizar muestreos rápidos y extensivos, no vemos ninguna objeción para que los datos así tratados puedan ser clasificados, sin olvidar que representan amplios intervalos.

Tabla 3.5. Escala de abundancias según Frontier (1969).

Clases	Intervalos de abundancias
1	de 1 a 3
2	de 4 a 18
3	de 18 a 80
4	de 80 a 350
5	de 350 a 1500

Datos cuantitativos

Los datos *cuantitativos*, siempre de multiestado, son los llamados cardinales, magnitudes o cantidades dado que miden relaciones cuantitativas en sentido estricto (Crisci y López Armengol, 1983). Su obtención involucra, por tanto, una medición a lo largo de una escala, y una unidad de medida (Jobson, 1991). Si la unidad de medida es indivisible se dice que la variable o el tipo de dato es *discontinuo* o *discreto* (Jobson, 1991). Los datos discontinuos solo pueden expresarse por números enteros de ahí que su variabilidad sea discontinua, como es el caso del número de especies (o de cualquier taxa), aunque si este parámetro se estima por métodos como el de rarefacción de Sanders (1968), el número de especies para un tamaño de muestra dado puede ser un número fraccionario, caso en el cual caería artificialmente en la categoría de continuo.

Si la unidad de medida es infinitamente divisible de modo que al menos teóricamente la medición puede hacerse en unidades cada vez más finas, se dice que el dato es *continuo* (Jobson, 1991). Los datos continuos expresan cualidades cuya variabilidad se distribuye en una escala continua, de ahí que su expresión pueda ser un número entero o fraccionario. Incluyen la mayoría de los parámetros ecológicos: abundancia, biomasa, cobertura, diversidad, dominancia, equitatividad; o la longitud, ancho o diámetro cuando se miden individuos de una población con propósitos de estudiar algún aspecto ecológico relacionado con su dinámica.

En el dato cuantitativo el atributo está representado por un número, que es expresión original de algún parámetro ecológico, bien sea estructural o funcional. En tales casos el análisis normal examinará cuánto se parecen las dos localidades en la composición, ya no solo cualitativa, sino también cuantitativa de sus especies; el análisis inverso, cuáles especies resultan más afines en sus patrones de distribución espacial o temporal.

Para la clasificación de los datos cualitativos aclarábamos que debían buscarse contrastes entre las entidades. Lo mismo es válido para los datos cuantitativos pero en este caso nos referimos a diferencias numéricas importantes que garanticen la formación de grupos, como se muestra en el ejemplo de la Tabla 3.6 que compara la abundancia de especies coralinas en ambientes limpios y contaminados (Herrera, 1991). Las estaciones limpias poseen una mayor representación de especies con dominancia de *Agaricia agaricites*, mientras que en las estaciones contaminadas el número de especies se reduce y dominan especies resistentes como *Siderastrea radians*. Ello se refleja en contrastes numéricos en la tabla que indican su idoneidad para la clasificación. Si los valores de la matriz son muy similares ello podría reflejar muestras procedentes de un biotopo muy homogéneo donde la creación de conjuntos puede tener poca significación ecológica y la afinidad varía en un estrecho intervalo. Si las diferencias numéricas entre los datos son de pequeña escala pero significativas para el agrupamiento del parámetro que se estudia deberán emplearse medidas de afinidad que ayuden a hacer evidentes dichas diferencias.

Distintos tipos de parámetros pueden brindar clasificaciones bien diferentes en lo cual influye, en primer lugar, su escala de variación. Por ejemplo, la heterogeneidad que se expresa en natios fluctúa entre 0 y 3; pero la cobertura porcentual entre 0 y 100 y la biomasa en gramos de peso húmedo

Tabla 3.6. Abundancia de especies de corales (colonias/m²) en nueve estaciones del arrecife costero al N de la Habana, Cuba, en un gradiente de contaminación al E y al O de la Bahía de la Habana y el Río Almendares, en comparación con áreas de referencia limpias en Playa Baracoa y Santa Cruz del Norte. Los subíndices del 1 al 4, indican el orden en el gradiente a partir de las fuentes contaminantes (según Herrera, 1991).

Especies	Playa Baracoa	Río Almendares			Bahía de la Habana				Sta. Cruz
		O ²	O ¹	E ¹	O ⁴	O ³	O ²	O ¹	
<i>Madracis decactis</i>	0.10	0.10	0	0.10	0	0	0	0.04	0
<i>Diploria labyrinthiformis</i>	0.10	0	0	0	0	0	0	0	0.02
<i>D. clivosa</i>	0	0	0	0	0	0	0	0	0.02
<i>D. strigosa</i>	0.05	0	0	0	0	0	0	0	0.02
<i>Manicina areolata</i>	0	0	0	0	0	0	0	0	0.02
<i>Montastraea cavernosa</i>	0.95	0.30	1.40	0.70	0.30	0.06	0	0.04	0.22
<i>M. annularis</i>	0.25	1.10	0.30	0	0	0	0	0	0.02
<i>Stephanocoenia michelini</i>	0.05	0	0.10	0.10	0	0.06	0	0.08	0
<i>Isophyllia sinuosa</i>	0	0	0	0	0	0	0	0	0.2
<i>Dichocoenia stokesi</i>	0.35	2.00	0.70	0.40	0.20	0.33	0	0.12	0.08
<i>Meandrina meandrites</i>	0.55	0.40	0.10	0	0	0	0	0	0.06
<i>Eusmilia fastigiata</i>	0.20	0	0.30	0	0	0	0	0	0
<i>Agaricia agaricites</i>	1.45	0.10	0	0.10	0	0	0	0.20	1.04
<i>Siderastraea radians</i>	1.00	0.80	2.50	3.15	3.30	1.06	2.10	1.56	0.12
<i>Porites porites</i>	0	0	0	0	0	0	0	0	0.04
<i>P. astreoides</i>	0.50	0.50	0.90	0	0	0	0	0	0.64
Densidad total	5.55	5.30	6.30	4.55	3.80	1.53	2.10	2.04	2.32

puede alcanzar valores mayores. Sin embargo, esta misma biomasa expresada en peso seco brindaría valores más moderados, con lo cual aclaramos que el cambio en la unidad de medida es también una fuente de variación. Este inconveniente se obvia seleccionando las variables más significativas a los efectos de la agrupación, estandarizándolas en la misma unidad (individuos/m² si se trata de abundancia, o gramos/m² si son datos de biomasa) para hacerlas comparables, o dándoles algún manejo estadístico.

Por otra parte, las diferencias entre clasificaciones efectuadas con distintos datos pueden tener una importante connotación teórica relacionada con lo que cada uno representa en la estructura o función de la comunidad que se estudia. Ignatiadis *et al.* (1992) encuentran resultados distintos al clasificar las comunidades fitoplanctónicas del Golfo de Saranicos (Mar Egeo) a partir de tres matrices: abundancia de especies (células/l); índices ecológicos (diversidad, equitatividad, dominancia, redundancia y riqueza); y parámetros relacionados con la biomasa (clorofila *a*, total de células, productividad primaria, tasa de asimilación y relación diatomeas/dinoflagelados).

La clasificación obtenida con abundancia ordena los grupos relacionados taxonómicamente y difiere de la obtenida con índices ecológicos, pues valores iguales de diversidad pueden corresponderse con composiciones de especies distintas. La discrepancia de estas clasificaciones con la de los parámetros relacionados con la biomasa ocurre pues aunque las muestras sean cuantitativamente similares en su estructura, la producción primaria o la tasa de asimilación están más influidos por el estado fisiológico de la célula y variables ambientales como la luz y la turbulencia (Ignatiadis *et al.*, 1992).

El empleo de datos cuantitativos requiere de un mayor análisis para indagar por el cumplimiento de algunos requisitos mínimos. Sin embargo, estos requisitos no se corresponden con los de homocedasticidad, normalidad o linealidad tan importantes en las técnicas de inferencia estadística pues el análisis de grupos es una metodología para agrupar observaciones y como tal tiene fuertes propiedades matemáticas pero ningún fundamento estadístico. Al respecto, según Hair *et al.* (1995) hay que considerar: la *representatividad* de la muestra y su *multicolinealidad*.

En referencia a la representatividad muestral, dado que el investigador cuenta con un censo de la población para obtener grupos y desea con éste lograr una estructura, debe tener seguridad de que la muestra obtenida es realmente representativa de dicha población y los resultados, por tanto, generalizables. Cuando se emplean unidades muestrales los estimados de cualquier parámetro ecológico, bien sea densidad, biomasa, cobertura u otro, deben provenir de un número suficiente de observaciones o réplicas que garanticen la confiabilidad individual de cada dato, en cada una de las localidades o intervalos de tiempo. Además, debe tenerse en cuenta la posible influencia de las variaciones estacionales naturales o las diferencias locales debido a irregularidades o cambios dentro del mismo hábitat que influyan en la distribución de las especies. El análisis de grupos será solo tan bueno como lo sea la representatividad de la muestra (Hair *et al.*, 1995).

En relación con la colinealidad aclaremos antes que este término define la fuerza de relación entre dos variables, que en el caso de tres o más se denomina multicolinealidad aunque a veces se emplea el término intercambiabilidad (Hair *et al.*, 1995). Dos variables serán completamente colineales si su coeficiente de correlación r es igual a 1, mientras que un r igual a 0 indicaría una falta total de colinealidad. Como en la medida en que dos entidades sean muy colineales su aporte a la afinidad -medida como correlación- será muy alta, la multicolinealidad impone la unión conjunta de un gran número de entidades pues éstas son fuertemente ponderadas (Hair *et al.*, 1995) y tienden a dominar en el agrupamiento (Krzanowski y Marriott, 1966a) decidiendo de antemano la solución de la clasificación. Este requisito cobra mayor importancia cuando trabajamos con coeficientes de correlación como medidas de afinidad pues las restantes medidas no responden de igual forma a la colinealidad. Por otra parte, no creemos que esta propiedad sea una regla de los datos ecológicos como sí lo es de algunas variables morfométricas o estructurales que emplea la taxonomía numérica, donde de hecho la falta de multicolinealidad es una propiedad deseable en los caracteres a seleccionar (Sneath y Sokal, 1973).

La categorización de los datos cuantitativos también puede hacerse sobre la base de las relaciones entre los elementos que componen su escala de medición, subdividiéndolos en datos de *escala de intervalo* y de *escala de relación*. Las variables que emplean estos tipos de escalas se denominan *métricas* (Everitt y Dunn, 1991; Jobson, 1991). En el dato basado en escala de intervalo las diferencias entre puntos sucesivos de la escala son iguales, pero el punto cero es arbitrario (Everitt y Dunn, 1991) pues no tienen un valor de base natural (Sharma, 1996). Jobson (1991) cita como ejemplo la temperatura donde la medición procede a partir de un punto fijo y la relación entre dos medidas en °C no se mantiene si lo convertimos en °F. Con este tipo de dato pueden emplearse todos los métodos estadísticos salvo los basados en datos de escala de relación. La escala de relación constituye

la más versátil y tiene como propiedades que: a) dos valores a lo largo de la escala pueden ser expresados significativamente como una relación; b) la distancia entre puntos de la escala es significativa; c) los elementos a lo largo de la escala pueden ser ordenados de menor a mayor (Jobson, 1991). Estos datos constituyen el nivel más acabado de medición, permiten hacer comparaciones entre las diferencias y sus magnitudes relativas (Everitt y Dunn, 1991) y no hay restricciones a las pruebas estadísticas a emplear (Sharma, 1996).

Reducción de los datos

Los muestreos ecológicos pueden generar grandes matrices de datos por el número de estaciones, si se trata de muestreos muy extensivos en el espacio, pero el mayor volumen de información se debe fundamentalmente a las largas listas de especies que típicamente presentan varias dominantes y una larga «cola» de otras relativamente raras. El número de especies que se registra durante un muestreo depende del grupo que se trate. En los arrecifes coralinos de la plataforma SO de Cuba por ejemplo, se pueden encontrar en un muestreo 21 especies de corales, 34 de gorgonáceos, más de 70 de esponjas, entre 60 y 80 de moluscos y muchas más en poliquetos o peces, por solo mencionar algunos grupos (ver Alcolado, 1990). Quiere esto decir que siempre es conveniente analizar si la matriz original de datos que va a ser sometida a clasificación debe ser reducida, lo cual de modo general se realiza eliminando o fusionando la información de varias estaciones, o eliminando algunas especies. Para esto último, un primer criterio elemental sería la eliminación de aquellas cuya identificación taxonómica sea dudosa.

De las tres razones que comenta Boesch (1977) por las cuales se hace necesario reducir los datos: a) disminuir el número de cálculos y por tanto el costo; b) poder emplear ciertas estrategias clasificatorias y c) excluir datos de poca significación, esta última, por su connotación ecológica, debe ser siempre analizada aún cuando se cuenten con todas las facilidades computacionales. En el patrón de distribución típico de la composición cuantitativa y cualitativa de especies en una comunidad siempre existen varias especies raras en pequeño número, cuya probabilidad de aparición es baja. Por esta razón sus relaciones de coocurrencia pueden deberse más al azar que a requerimientos similares de hábitat por lo cual no brindan patrones precisos de distribución, al menos dentro de los marcos de un esfuerzo de muestreo razonable. Ello avala el criterio más usado para eliminar especies de una matriz: su baja frecuencia dentro de la muestra.

Otros criterios empleados han sido la eliminación de especies con poca contribución a la varianza cada cierto número de muestras (Bakus, 1990) o la exclusión de especies cuya abundancia o constancia caiga por debajo de determinado valor dentro de la información total de las localidades o intervalos de muestreo (Boesch, 1977). Aunque se han ensayado métodos más complicados, consideramos que los criterios de frecuencia, constancia y abundancia debidamente analizados en la matriz original de datos, pueden conducir al investigador que domina la ecología de su grupo a realizar las reducciones necesarias de especies.

No debe olvidarse que existen especies restringidas a determinados hábitats. Por ejemplo, de las varias zonas del arrecife de barrera el octocoralio *Gorgonia flabellum* se encuentra representada casi exclusivamente en la zona de embate por su alta resistencia al batimiento y puede estar ausente

o tener escasa abundancia en las restantes zonas. Considerar aisladamente su frecuencia o constancia respecto a todo el arrecife podría hacer pensar en su eliminación pero si consideramos su abundancia unido a los criterios sobre su zonación ecológica, podemos discernir que la misma es clave para la clasificación de la zona arrecifal pues tipifica su región más expuesta.

En relación con la reducción de la matriz eliminando localidades o intervalos de tiempo, Boesch (1977) señala como criterios más empleados la eliminación de muestras de dudosa calidad y la fusión de varias estaciones para lograr mayor homogeneidad. A esto último podríamos añadir que la unión de varias estaciones podría ser también una variante para incrementar el tamaño de muestra aunque tengamos que analizar la comunidad en un mayor contexto espacial. Por ejemplo, varias estaciones de una pradera marina de *Thalassia testudinum* pueden estar submuestreadas individualmente pero en conjunto brindar un panorama claro de la composición del biotopo, siempre y cuando no existan diferencias importantes en sus características sedimentológicas e hidrológicas.

También pueden combinarse muestras de distintas épocas si se desea dilucidar los patrones espaciales sin contar con la información temporal, o por el contrario, combinar muestras de distintas localidades para analizar los patrones temporales. La reducción de datos no se practica, cuando tratamos con datos cualitativos donde todos tienen el mismo “valor” pero puede ser un paso necesario con los cuantitativos aunque advertimos del peligro que una reducción excesiva puede conducir a obtener clasificaciones artificialmente precisas.

Transformación de los datos

La transformación de los datos es un paso conocido de la estadística clásica cuando se requiere la normalización de los valores originales para poder aplicar determinadas pruebas paramétricas. Por definición, la *transformación* es una alteración del valor del atributo sin referencia al intervalo de valores de la población como un todo, o sea, que concierne a cada atributo individualmente y no a la totalidad de la muestra (Boesch, 1977). Al transformar se pretende ante todo reducir la importancia excesiva de las especies más abundantes o corregir la ausencia de éstas (Bakus, 1990). Una de las transformaciones más aplicadas es la logarítmica, bien sea en su expresión $\log X$, o $\log (X+1)$ si existen ceros en la matriz de datos; otras también comunes son la de raíz cuadrada, cúbica y arcoseno.

Pero el dato original también puede ser modificado mediante *estandarización*, que se diferencia de la transformación propiamente dicha en que la alteración de los datos sí depende de alguna propiedad del ordenamiento de los valores como un todo. Son ejemplo de ella la conversión de los valores en porcentajes o proporciones, la expresión de los valores en unidades de desviación estándar y la estandarización doble. Al estandarizar se pretende reducir los valores a una escala común de comparación ante grupos de diferentes tamaños o grados (Bakus, 1990).

Boesch (1977) explica que se hace necesario transformar los datos cuando: a) existen grandes diferencias entre los valores, b) su distribución se aleja de la normalidad o c) el esfuerzo de muestreo no ha sido uniforme. Frecuentemente, en los datos cuantitativos unas pocas especies tienen valores

excesivamente altos y el resto muy bajos. Como algunos índices de afinidad son muy sesgados a los altos valores, de modo que las especies abundantes determinan el valor de afinidad, puede ser aconsejable una transformación (Fig. 3.4) que atenúe su contribución.

Estaciones						Estaciones					
Especies	1	2	3	4	5	Especies	1	2	3	4	5
A	123	982	478	50	27	A	4.9	9.9	7.8	3.7	3.0
B	67	10	13	1	0	B	4.1	2.1	2.3	1.0	0
C	15	35	234	5	0	C	2.5	3.3	6.2	1.7	0
D	2	1	3	1	0	D	1.2	1.0	1.4	1.0	0
E	1	0	1	79	8	E	1.0	0	1.0	4.3	2.0

Figura 3.4. Matriz original de datos cuantitativos y transformada mediante raíz cúbica.

Por otra parte algunos coeficientes derivados de la estadística, como la correlación, pueden emplearse como medidas de afinidad pero para ello se requiere que los datos se normalicen mediante alguna transformación. Finalmente, cuando las diferencias en el esfuerzo de muestreo entre dos localidades no permite la comparación directa de los datos, como ocurre por ejemplo con muestreos del bentos por recorrido (espacial o temporal) o con equipos de rastreo, es conveniente la estandarización respecto al total de estaciones, donde las abundancias de cada especie en una estación se suman y se dividen entre el total, con lo cual los datos quedan expresados en forma de proporciones, o porcentajes si se multiplican por 100 (Fig. 3.5).

De igual forma puede realizarse la estandarización por total de especies sumando las abundancias de cada una en todas las estaciones y dividiendo entre el total. Esta estandarización puede ponderar fuertemente las especies raras y subestimar las comunes por lo que se recomienda solo si la frecuencia de especies en la tabla no son muy diferentes (van Tongeren, 1987). Cuando estandarizamos por estaciones los porcentajes «valen» en el sentido de las proporciones de especies en la localidad (o sea por columnas), lo cual tiene por supuesto un significado ecológico.

Estaciones						Estaciones					
Especies	1	2	3	4	5	Especies	1	2	3	4	5
A	120	98	82	3	1	A	64.9	76.5	49.7	50.0	100
B	57	25	41	0	0	B	30.8	19.5	24.8	0	0
C	5	3	32	0	0	C	2.7	2.3	19.4	0	0
D	2	1	10	0	0	D	1.1	0.8	6.1	0	0
E	1	1	0	3	0	E	0.5	0.8	0	50.0	0

Figura 3.5. Matriz original de datos cuantitativos y matriz estandarizada por estaciones mediante porcentajes.

Los datos de rastreos o recorridos en distintas estaciones no deben ser estandarizados en porcentajes por especies (por filas) si los esfuerzos de muestreo son diferentes. Una estandarización porcentual por filas podría justificarse solo si los datos están referidos a la misma unidad. De cualquier manera, para estandarizar los datos mediante porcentajes o proporciones ambas localidades deben tener un

tamaño de muestra lo suficiente elevado que aproxime adecuadamente las proporciones calculadas. En el ejemplo de la Figura 3.5 el tamaño de muestra de las estaciones 4 y 5 no es adecuado para que sea válida esta estandarización.

Existen otros tipos señalados por Boesch (1977) como la estandarización doble simultánea que resuelve los problemas de escala en el análisis normal e inverso; y las estandarizaciones por máximo de estaciones y máximo de especies (van Tongeren, 1987), donde los valores individuales se dividen respectivamente entre el valor de la especie con abundancia máxima en una estación, o entre la abundancia máxima que alcanza una especie en todas las estaciones en que aparece.

Tratamiento de datos mezclados

Aunque a lo largo de este texto veremos matrices de estaciones/especies con datos de igual categoría, vamos a referirnos brevemente a cuando ocurren mezclas de varios tipos de datos en la matriz original para que el lector se sienta orientado para manejar su información adecuadamente. El tratamiento de datos mezclados puede variar según la disciplina que se trate pero Kaufman y Rousseeuw (1990) lo resumen en tres posibilidades.

Una primera y lógica aproximación es separarlos y realizar clasificaciones independientes para los tipos compatibles y relacionar después sus resultados. Si estos dan muy diferente puede ser más práctico la segunda: procesarlos juntos tratando todas las variables como cuantitativas, o por el contrario, y como tercera variante, reducirlo todo a datos cualitativos. Esto último tiene la ventaja de la sencillez pero la desventaja de sacrificar información potencialmente útil (Everitt, 1993).

Veamos esto con un ejemplo tomando una selección de datos de diferentes categorías de nuestro estudio de la langosta *Panulirus argus* y sus refugios naturales en los arrecifes del Suroccidental de Cuba (Tabla 3.7) cuyo objetivo era definir las regularidades de su conducta selectiva (Herrera *et al.*, 1991). Como datos cualitativos de estados excluyentes tenemos el sexo de la langosta refugiada: hembra o macho; y de presencia-ausencia la naturaleza del sustrato en la base del refugio, según se presente o no sedimento arenoso. El tipo de refugio: cuevas, huecos y solapas, representa al dato cualitativo de multiestado desordenado donde no existe una jerarquía y se convierten por tanto en datos de presencia-ausencia. El estadio de muda, proceso fisiológico que al transitar por una etapa

Tabla 3.7. Algunos parámetros cuantitativos y cualitativos de la langosta *Panulirus argus* y sus refugios en los arrecifes de la plataforma Suroccidental de Cuba.

Refugio	Parámetros de la langosta			Parámetros del refugio			
	Sexo	Longitud total (cm)	Estadio de muda	Profundidad del refugio (cm)	Número de entradas	Sedimento en la base	Tipo de refugio
1	Macho	30.2	Intermuda	57.6	1	Si	Cueva
2	Hembra	48.6	Postmuda	23.1	1	No	Hueco
3	Macho	21.8	Postmuda	11.5	1	Si	Solapa
4	Hembra	29.4	Postmuda	12.8	2	Si	Hueco
5	Hembra	45.2	Premuda	20.5	1	No	Cueva

de pre, inter y postmuda permite un orden jerárquico asignando números del 1 al 3; ejemplifica el dato de multiestado ordenado. Como datos cuantitativos continuos tenemos la longitud total de la langosta (LT) y la profundidad de su refugio (PR), parámetros cuya relación (PR/LT) se mantiene cercana a 2; y como ejemplo de datos cuantitativos discontinuos el número de entradas del refugio, generalmente 1 y ocasionalmente 2 aberturas.

En la Tabla 3.8 hemos tratado toda la información como datos cualitativos. Los datos cualitativos originales han sido codificados de acuerdo a su categoría pero los cuantitativos han sido convertidos en los primeros. Una variante de codificación podría ser la de convertir los datos cuantitativos continuos en datos de multiestado ordenado codificando las tallas en grandes (1), medianas (2) y pequeñas (3) según un intervalo previamente establecido y lo mismo para la profundidad del refugio. Sin embargo como en el proceso de conversión es importante mantener las relaciones como base de su posterior clasificación resulta más adecuado en este caso agruparlos en intervalos para convertirlos en datos de presencia-ausencia de modo que profundidades del refugio menores de 20 cm se corresponden en la tabla con longitudes totales menores de 30 cm, y viceversa.

Tabla 3.8 Conversión de la información de la Tabla 3.7 en datos cualitativos.

Refugio	Sexo		LT (cm)		Estado de muda	Profundidad del refugio (cm)		Número de entradas		Sedimento en la base		Tipo de refugio		
	Macho	Hembra	< 30	> 30		< 20	> 20	1	2	Si	No	Cueva	Hueco	Solapa
1	1	0	0	1	2	0	1	1	0	1	0	1	0	0
2	0	1	0	1	3	0	1	1	0	0	1	0	1	0
3	1	0	1	0	3	1	0	1	0	1	0	0	0	1
4	0	1	1	0	3	1	0	0	1	1	0	0	1	0
5	0	1	0	1	1	0	1	1	0	0	1	1	0	0

En la Tabla 3.9 la información ha sido tratada cuantitativamente. Los datos cualitativos originales aportan el valor proveniente de su codificación bien sea 0 ó 1, ó entre 1 y 3; pues no tienen otro contenido de información. Los datos cuantitativos mantienen sus valores originales aunque alternativamente, para aliviar diferencias escalares pueden convertirse en datos semicuantitativos de rango, asignando números del 1 en adelante. De esta forma tenemos “números” en toda la tabla y la información se trata posteriormente como cualquier dato cuantitativo, incluido si necesitan alguna transformación (Crisci y López Armengol, 1983).

Tabla 3.9 Conversión de la información de la Tabla 3.7 en datos cuantitativos.

Refugio	Sexo		LT (cm)	Estado de muda	Profundidad del refugio (cm)	Número de entradas	Sedimento		Tipo de refugio		
	Macho	Hembra					Si	No	Cueva	Hueco	Solapa
1	1	0	30.2	2	57.6	1	1	0	1	0	0
2	0	1	48.6	3	23.1	1	0	1	0	1	0
3	1	0	21.8	3	11.5	1	1	0	0	0	1
4	0	1	29.4	3	12.8	2	1	0	0	1	0
5	0	1	45.2	1	20.5	1	0	1	1	0	0

En este proceso de conversión los seis parámetros originales (Tabla 3.7) quedan convertidos en doce (Tabla 3.8) y once (Tabla 3.9) para los datos cualitativos y cuantitativos, respectivamente. En este incremento influyen sobre todo los datos cualitativos de multiestado ordenado que deben ser codificados individualmente como presencia-ausencia provocando una sobreestimación del parámetro “tipo de refugio”.

Con este ejemplo sencillo solo pretendemos proveer al interesado con una pauta de trabajo, demostrar la libertad que se debe tener en el manejo de la información y la importancia de conocer los tipos de datos y sus relaciones. Ejemplos similares los brindan Crisci y Armengol (1983) para la taxonomía numérica.